NERU: Named Entity Recognition for German

Daniel Weber, Josef Pötzl
CIS, Ludwig Maximilian University, Munich
forwarding@go4more.de
j.poetzl@campus.lmu.de

Abstract

In this paper [1], we present our Named Entity Recognition (NER) system for German – NERU (Named Entity Rules), which heavily relies on handcrafted rules as well as information gained from a cascade of existing external NER tools. The system combines large gazetteer lists, information obtained by comparison of different automatic translations and POS taggers. With NERU, we were able to achieve a score of 73.26% on the development set provided by the GermEval 2014 Named Entity Recognition Shared Task for German.

1 Introduction

Generally, named entities (NEs) are phrases that represent persons, organizations, locations, dates, etc. For example, the German sentence "*Frau Maier hat einen Toyota aus Amerika gekauft.*" contains three named entities *Frau Maier*, which refers to a person, *Toyota*, referring to an organization and *Amerika*, marking a location. Embedded NEs may also be present, for example: *Troia - Traum und Wirklichkeit* is a NE, which contains an embedded NE of type location – *Troia*.

In this paper, we describe NERU, which is a rule-based system for NER for German that was developed in the context of the GermEval 2014 NER Shared Task that specifically targets only this language. Thus, NERU aims to identify not only flat NE structures, but as well embedded ones. As described by Benikova et al. (2014b), the maximal level of embedding for the GermEval 2014 task is one named entity. The main targeted types are PER (person), LOC (location), ORG (organization) and OTH (other) with two possible subtypes relevant for all four groups – deriv and part. The latter leads to a combination of 12 different NE types.

Following, in section 2, we discuss the motivation behind GermEval 2014 and the state-of-the-art approaches to NER focusing on the language important for this task – German. Then, in section 3, we provide more details on the structure of NERU and the approach we used. In section 4, we present the performance of the system on the development data provided by the GermEval 2014 shared task. Finally, in section 5, we conclude our work.

2 Related Work

NER is an important subtask of a wide range of Natural Language Processing (NLP) tasks from information extraction to machine translation and often even requires special treatment within them (Nagy T. et al., 2011). GermEval's goal is, however, to consider NER proper and to advance the state-of-the-art of this task for a particular language – German. This language has been rarely the focus within previous NER research, which mostly explores English. The CoNLL-2003 Shared Task on Language-Independent NER (Tjong Kim Sang and De Meulder, 2003) addressed this problem and included German as one of its targets, although, in general, multilin-

guality was the objective.

While the majority of NER so far was concentrating almost only on flat NE structures (Tjong Kim Sang and De Meulder, 2003; Finkel and Manning, 2009), one of the main goals of GermEval is also to push the field of NER towards nested representations of NEs. Independent of the NE representation itself, there are many different approaches to tackle this task, for example, by using machine-learning techniques, such as Hidden-Markov-Models (Morwal et al., 2012), rule-based (Riaz, 2010) or even a combination of both (Nikoulina et al., 2012). NE recognition utilizing a hybrid approach has also been performed by Saha et al. (2008), who presented a set of handcrafted rules together with the use of gazetteer lists which were transliterated from English to Hindi with their own transliteration module.

As German significantly differs from other languages regarding capitalization or syntax in general, some of the common approaches, specifically on English, can not be transferred to German automatically. Thus, in the context of GermEval, we concentrate mostly on handcrafted rules as well as information from external NER tools. The full pipeline of the NERU system is presented in more detail further in section 3.

3 The NERU System

NERU's pipeline is structured as follows: In a first step, we use vast gazetteer lists to attain first suggestions for NEs (see section 3.1). Secondly, we utilize automatic translation tools to find matches occurring in various languages (described in section 3.2). Thirdly, the results of the TreeTagger (see section 3.3), the Stanford NE Recognizer (see section 3.4) and examining contexts of NE's (see section 3.5) are then taken into consideration. The combination results to a cascade of different methods that provide a set of suggestions for the NEs in the data. In a last step, we revise this set and modify it by removing and altering its entries with a number of manually crafted rules (see section 3.6).

3.1 Gazetteers

Gazetteers are predefined word lists which represent standard sources for NER as they contain NEs, such as names, organizations and locations marked for their correct category. So far, gazetteers were widely employed for tackling this task (Kazama and Torisawa, 2008; Jahangir et al., 2012; Alotaibi and Lee, 2013). NERU also employs gazetteers (mainly lists of locations and persons), which were collected from the German Wikipedia[2] and then manually extended.

One of the biggest problems in NER is resolving ambiguity. If all NEs are unambiguously identifiable, a large gazetteer would be sufficient. In natural language, however, there are context-sensitive terms, such as *California Institute of Technology*, which can on the one hand appear as a location and on the other as an organization. The decision as to which category the Named Entity shall be assigned depends solely on its textual environment.

3.2 Preclusion Through Translation

To deal with false-positives generated with the use of gazetteers, more sophisticated methods are needed to perform viable NER. In order to also consider the textual environment of the tokens, we make use of machine translation (MT). In fact, translations of NEs often leads to the use of the same surface form in both languages, specifically most proper names are not affected by the translation procedure. Therefore, we assume that all tokens that do not change within translation are reasonable NE candidates.

The Google Translate API[3] is used for translating the German data into English. For stopwords that are present in both languages, which should not be marked as NEs, we incorporated a list created by the intersection of the lists of stopwords from both English and German.

3.3 TreeTagger

To provide further suggestions for NEs, we employ the TreeTagger (Schmid, 1994; Schmid, 1999), which is a robust POS tagger for German reaching state-of-the-art performance. The tagger may also be partially used as a recognizer when the POS tags for proper names (*NE*) are employed. Hence, all tokens tagged with the *NE* tag are also considered as NE candidates.

[2] https://de.wikipedia.org

[3] https://developers.google.com/translate

3.4 Stanford NER

In the search for a wider source of diverse suggestions for the NEs in the data, we embedded the Stanford NER[4] in our system to find additional candidates for NEs. It is very robust in detecting NEs, however being restricted to only one type of NE – PER. All tokens marked as NE by the Stanford NER are again used as NE candidates by NERU.

3.5 Context Frequency and Probability

Using the GermEval training data, we also detect potential NEs by observing their type and frequency of contexts. If token t is marked by a NE tag (e.g. B-LOC, I-PER, etc.), we extract a NE-trigram (t_{-1}, t, t_{+1}) for it. Frequency counts of the trigrams are then collected and the ones occurring less than 5 times are ignored. Following, the probability of a token in a specific context is calculated. Only tokens that have a probability > 0.5 of being in that context are marked as NEs.

Assuming a token sequence "*der philippinischen Hauptstadt*" is encountered, "*philippinischen*" would be tagged as B-LOCderiv. If there are different options for a NE tag in this context, the option with the highest probability is chosen.

3.6 Rule-Based Filtering

In sections 3.1 through 3.4, we presented a number of different approaches, which we used for the identification of NEs in the data. This cascade of modules, however, results to a generously tagged dataset including suggestions for as many NEs as possible. In order to reduce this set, in the last step of NERU's pipeline, we process the output with the help of a collection of handcrafted rules. An additional set of rules is also used that relies only on the information provided by the gazetteers and manually created lists of abbreviations.

3.7 Rules for Person NEs

To identify NE of the type PER, we examine contexts and tokens we categorized as trigger words, such as nobiliary particles, honorary or heredity titles, etc. For example, Roman numerals may indicate a person (e.g. *Karl IV*), similar to the generational title "*Jr.*", which may also appear following the candidate NE. Additionally, when particles, such as "*von*" or "*de*" are found between two or more NEs of the type PER or the special case that a NE of the type LOC is perceived right after "*von*" (*of*), the latter are combined to one single span, for example "*Wilhelm Friedrich Ludwig von Preußen*".

[4] http://nlp.stanford.edu/software/CRF-NER.shtml

3.8 Rules for Organization NEs

For the identification of organizations, we looked for special characters like "*&*" between NEs of type PER (e.g. *Kanzlei Heigl & Hartmann*). We furthermore deduce organization names from common abbreviations. If a token is found, which is marked as a LOC or a PER and its preceding token is a common abbreviation (e.g. *AC*, *TSV* etc., which we check against a manually created list of common abbreviations), then the whole sequence indicates a NE of type ORG (e.g. *FC Barcelona*).

In a similar way, the abbreviations for a type of organization, such as "*GmbH*", "*Comp.*", "*KG*" are also used as indicators for NEs of type ORG. Such tokens or their attributed NEs are combined with any closely preceding NE of type ORG or PER. It is not distinguished between the types ORG and PER, as we consider organization names like "*Wortmann AG*". We investigate the preceding tokens until a token which has been tagged as ORG or PER is found, unless the examined sequence is larger than 5 tokens. In this case, the 5th token is chosen automatically. For example, if "*Bandidos Kapital und Invest AG*", is considered and only the token "*Bandidos*" is already tagged as a NE of type ORG, the identification of the abbreviation "*AG*" would impose the marking of the full span as NE of type ORG.

3.9 Rules for Location NEs

In order to recognize location names, we look for specific character patterns, such as "*straße*" (street) in the tokens (e.g. *Leopoldstraße*). Once more, we investigated the contexts to properly find connected sequences. For example, when a number is preceded by a NE of type LOC, the number is also included into the NE sequence (e.g. "*Dachauer Straße 24*").

setting	strict				loose				outer				inner			
	Acc.	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**	**Acc.**	**P**	**R**	**F1**
CF	93.58	15.32	10.67	12.58	93.59	15.76	10.98	12.95	87.73	15.32	11.52	13.15	99.42	0.00	0.00	0.00
TT	95.34	28.98	14.45	19.28	95.35	29.26	14.59	19.47	91.26	28.98	15.59	20.28	99.42	0.00	0.00	0.00
St	95.81	70.34	15.04	24.78	95.81	70.34	15.04	24.78	92.20	70.34	16.23	26.37	99.42	0.00	0.00	0.00
Rul	98.19	72.30	74.26	**73.26**	98.28	74.60	76.61	**75.59**	96.93	72.92	78.05	**75.40**	99.45	54.90	26.42	**35.67**
St/TT	96.20	51.93	29.31	37.48	96.20	52.18	29.45	37.65	92.97	51.93	31.64	39.32	99.42	0.00	0.00	0.00
$St/TT/CF$	94.59	28.71	33.30	30.84	94.61	29.10	33.75	31.25	89.77	28.7	35.94	31.92	99.42	0.00	0.00	0.00
$St/TT/Rul$	98.01	67.52	74.91	71.02	98.11	69.61	77.23	73.23	96.58	67.94	78.76	72.95	99.45	54.90	26.42	**35.67**
all	96.28	46.07	75.02	57.09	96.37	47.50	77.34	58.85	93.11	45.88	78.8	58.01	99.45	54.90	26.42	**35.67**

Table 1: Results achieved by NERU based on the GermEval development set.

4 Evaluation

The evaluation of the progam will be done by the standard precision, recall and F1 score metrcis and some enhanced metrics, which is used to determine the overall ranking of the system. [5]

NERU was evaluated on the GermEval development set. We tested a number of settings: CF – tagging the data only based on the probabilities calculated on the context frequencies, TT – tagging the data only based on TreeTagger's POS tags, St – using only the Stanford NER, Rul – employing only the handcrafted rules. Further, combinations of these settings are also tested. In table 1, we list the respective system scores.

Considering the results on the strict evaluation setting, NER based only on context probabilities (CF) achieves 12.58%, which is the lowest performing setting of the system, followed by the use of the TreeTagger (TT) with 19.28% and the Stanford NER (St) with 24.78%. Surprisingly, NERU's best performance (73.26%) is reached only via the use of handcrafted rules (Rul), where all external tools (TreeTagger and Stanford NER) are not used. Using the information provided by the latter leads to a decrease of system performance to 71.02% ($St/TT/Rul$). This is a somewhat surprising result, considering the fact that the TreeTagger and the Stanford NER identify a significant portion of the NEs on their own (St/TT) reaching a score of 37.48%. Our assumption, however, is that this additional information contradicts the conclusions met by the rules that are solely based on gazetteers and abbreviation lists, which also leads to the decrease of scores. Thus, the final version of the system that we used for the annotation of the GermEval test set employs only the system setting Rul.

Looking deeper into this system setting (based on the system scores presented in table 2), we can see that NERU does not tag at all a large portion of the NE subtypes: *ORGderiv, OTHderiv, OTHpart, PERderiv, PERpart*. After qualitatively evaluating a sample of the system output, we could see that most of these subtypes are generally marked as their supertypes, e.g. *ORGderiv* is tagged as *ORG*. Another observation we could make on this sample is the fact that NERU tends to overgenerate and mark a good portion of non-NE tokens as NEs, e.g. *Bundeswehr*, *Waffen-SS* or *Bundesliga*.

[5] GermEval 2014 NER Evaluation plan `http://is.gd/eval2014`

NE Typ	Precision	Recall	FB1
LOC	84.42%	85.14%	84.78
LOCderiv	88.28%	89.79%	89.03
LOCpart	92.11%	67.31%	77.78
ORG	54.69%	69.15%	61.08
ORGderiv	0.00%	0.00%	0.00
ORGpart	96.55%	92.31%	94.38
OTH	61.27%	57.43%	59.28
OTHderiv	0.00%	0.00%	0.00
OTHpart	0.00%	0.00%	0.00
PER	75.89%	87.41%	81.25
PERderiv	0.00%	0.00%	0.00
PERpart	0.00%	0.00%	0.00
Strict			73.26

Table 2: Detailed scores on the strict evaluation setting based on the Rul system setting.

4.1 Offical Score

Regarding the offical score (Benikova et al., 2014a) NERU lost 25 % of performace in comparison with the development set. The system reached an accuracy of 96.96, a precision of 62.57, a recall of 48.35 and a resulting F_1 of 54.55 in the test set run. The score was calculated by the offical metrics used for the GermEval 2014 Shared Task. An explanation of this losses could be that NERU was also trained with the develop-

Metric	Acc.	P	R	F1
strict	96.96	62.57	48.35	54.55
loose	97.00	63.62	49.16	55.46
outer	94.56	63.69	51.33	56.84
inner	99.37	33.85	12.62	18.39

Table 3: Offical results on test set for all metrics.

ment set in some special cases. Also, as previously mentioned, we did not tag all Named Entity subtypes (6 ouf of 12 types are not taken into consideration).

5 Conclusion

The current paper presents the NER system NERU, which makes use of handcrafted rules, gazetteers and external NER tools for the recognition of NEs in the data. We evaluated the system on the GermEval development set, which showed that the handcrafted rules that do not use the information provided by the TreeTagger and the Stanford NER reach optimal system performance. These rules are solely based on gazetteers and manually created abbreviation lists. Using the latter, NERU participated in the GermEval 2014 NER Shared Task reaching 73.26% on the strict evaluation setting, which is a considerably good performance for German with respect to the scores reported for this language during the CoNLL-2003 Shared Task.

References

Fahd Alotaibi and Mark Lee. 2013. Automatically Developing a Fine-grained Arabic Named Entity Corpus and Gazetteer by utilizing Wikipedia. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 392–400, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Darina Benikova, Chris Biemann, Max Kisselew, and Sebastian Pado. 2014a. Germeval 2014 named entity recognition: Companion paper. In *Proceedings of the KONVENS GermEval Shared Task on Named Entity Recognition*, Hildesheim, Germany.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014b. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Jenny Rose Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 141–150, Singapore, August. Association for Computational Linguistics.

Faryal Jahangir, Waqas Anwar, Usama Ijaz Bajwa, and Xuan Wang. 2012. N-gram and Gazetteer List Based Named Entity Recognition for Urdu: A Scarce Resourced Language. In *Proceedings of the 10th Workshop on Asian Language Resources*, pages 95–104, Mumbai, India, December. The COLING 2012 Organizing Committee.

Jun'ichi Kazama and Kentaro Torisawa. 2008. Inducing Gazetteers for Named Entity Recognition by Large-Scale Clustering of Dependency Relations. In *Proceedings of ACL-08: HLT*, pages 407–415, Columbus, Ohio, June. Association for Computational Linguistics.

Sudha Morwal, Nusrat Jahan, and Deepti Chopra. 2012. Named Entity Recognition using Hidden Markov Model (HMM). In *International Journal on Natural Language Computing (IJNLC)*, volume 1.

István Nagy T., Gábor Berend, and Veronika Vincze. 2011. Noun Compound and Named Entity Recognition and their Usability in Keyphrase Extraction. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 162–169, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Vassilina Nikoulina, Agnes Sandor, and Marc Dymetman. 2012. Hybrid Adaptation of Named Entity Recognition for Statistical Machine Translation. In *Proceedings of the Second Workshop on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid MT*, pages 1–16, Mumbai, India, December. The COLING 2012 Organizing Committee.

Kashif Riaz. 2010. Rule-Based Named Entity Recognition in Urdu. In *Proceedings of the 2010 Named Entities Workshop*, pages 126–135, Uppsala, Sweden, July. Association for Computational Linguistics.

Sujan Kumar Saha, Sudeshna Sarkar, and Pabitra Mitra. 2008. A Hybrid Feature Set based Maximum Entropy Hindi Named Entity Recognition. In *IJCNLP*, pages 343–349.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International*

Conference on New Methods in Language Processing, pages 44–49, Manchester, UK.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Processing*, pages 13–26. Kluwer Academic Publishers, Dordrecht.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Walter Daelemans and Miles Osborne, editors, *Proceedings of CoNLL-2003*, pages 142–147. Edmonton, Canada.